

Short-term load forecasting with using multiple linear regression

Bhatti Dhaval, Anuradha Deshpande

Department of Electrical Engineering, Faculty of Technology and Engineering,
Maharaja Sayajirao University of Baroda, India

Article Info

Article history:

Received Jun 25, 2019

Revised Feb 25, 2020

Accepted Mar 3, 2020

Keywords:

Multiple linear regression

Regression

Short term load forecasting

ABSTRACT

In this paper short term load forecasting (STLF) is done with using multiple linear regression (MLR). A day ahead load forecasting is obtained in this paper. Regression coefficients were found out with the help of method of least square estimation. Load in electrical power system is dependent on temperature, due point and seasons and also load has correlation to the previous load consumption (Historical data). So the input variables are temperature, due point, load of prior day, hours, and load of prior week. To validate the model or check the accuracy of the model mean absolute percentage error is used and R squared is checked which is shown in result section. Using day ahead forecasted data weekly forecast is also obtained.

Copyright © 2020 Institute of Advanced Engineering and Science.
All rights reserved.

Corresponding Author:

Anuradha Deshpande,
Department of Electrical and Engineering,
Maharaja Sayajirao University of Baroda,
Kala bhavan, Near BadamadiBaug, Vadodara, Gujarat, India.
Email: a.s.deshpande-eed@msubaroda.ac.in

1. INTRODUCTION

In today's world continuity in electrical supply is necessary. Electricity is an essential need in daily life of people. Electrical utilities try to provide continuous power supply to their customers. To ensure this load forecasting is an important aspect. Load forecasting is about estimating future consumptions based on various data and information available and as per consumer behavior. Load forecasting has an effective role in economic operation of power utilities. Short-term load forecasting refers to forecasting load for hour, day or week ahead. There are many tools and techniques available for load forecasting such as simple time series, regression, neural networks etc. Load forecasting mean forecasting average load in KW or total load in KWh for periods or blocks of 15 minutes, 30 minutes, 1 hour, day, week, month or a year for daily forecast, weekly forecast, monthly forecast or yearly.

There are many factors [1] which influence the accuracy of load forecasting like weather variables, holidays, festivals or events, tariff structures, available historical data, time of the year, day of the week and hour of the day. Weather variable includes temperature, humidity, rain and wind. Temperature and humidity has a considerable effect on power consumption because as the temperature rises people will turn on air conditioners and if temperature is low then air heaters will get turned on, which causes increase in electricity demand. If there is festival, the electricity demand will rise due to lightings.

Data used for STLF is mostly historical hourly load and weather parameters. Hourly load data is subject to errors made by instruments and/or operators and also unwanted or unplanned changes in the power system. Paper has considered five different cases of days and Absolute normalized residual method is used for improper data marking. Absolute normalized residual method is used for identification and modification of improper data [2].

Good quality of data is essential in power system before taking decision. So data must be cleaned and filtered before operator takes any decision from the data. Otherwise it will cause hazardous condition. Poor quality of data affects decision making without knowledge of operator. Data smoothing is necessary for some application for e.g. load forecasting in power system. Techniques used for this are Statistical techniques such as OWA (optimally weighted average) and MA (moving average). Data smoothing and filtering is done on smart meter data with measurements of 15 minute interval of DONWOD region of New York [3].

In Paper [4] Short term load forecast will not assume the contribution of agricultural load which is season dependent and thus is assumed to be constant in the span of few days. The triangular membership function is used where the support of the membership function is decided on the basis of the collected data. Relationship of load with weather variables, temperature and humidity is shown. These parameters are modeled in the domain of fuzzy, to obtain realistic value of the forecasted load when compared with the actual load. In paper [5] ANN (artificial neural network) technique is used for STLF. Three models are considered, 1. Prediction of the next hour load, 2. Prediction of the next day load profile, and 3. Prediction of the next day peak load. ANN models uses load profile and weather situation as input layer and results in forecasted next hour load, next day load profile, and next day peak load. Mean absolute error is used to evaluate the performance of Forecasting.

In paper [6] multiple linear regression analysis technique is used to forecast damping in Nordiac power system. A static MLR model is developed to explain the variability of the damping of the 0.35-Hz inter-area mode in the Nordic system. In [7] five short term load forecasting techniques are discussed. These five techniques are (i) Multiple linear regression, (ii) Stochastic time series, (iii) General exponential smoothing, (iv) State space and Kalman filter, and (v) Knowledge-based approach. Algorithms and necessary equations for these techniques are discussed. For other parameters, methods of short term load forecasting, controlling factors etc, discussed in [8-18]. In comparison to [19-29], proposed method calculates the short term load forecast using multiple regression method by calculating the random error (5%) which is showing the difference between observed and fitted model. The trained model is 95% accurate. Also the effect of variables is included.

2. RESEARCH METHOD

Method used for Short-term Load Forecasting is multiple linear regression. Regression analysis helps to know the relationship between load and variables affecting load consumptions like weather variable, seasonal effects and previous load consumption data. This method is based on statistical approach. Regression analysis helps to find out unknown values of parameters on the basis of given data set (available historical data). The calculation of unknown values of parameters can be explained by taking simple regression model. Simple linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (1)$$

Here, $i = 1, 2, n$

y is electrical load

β_0 and β_1 are Parameters (Regression coefficients)

x is explanatory variable

ε_i is random error

In this model relationship between load y and variable x is expressed with the help of parameters. We can obtain the value of load y for known value of variable x , if we have the β values of parameters. So, if we know the β values of parameters, we can obtain the values of load at any given values of x . ε_i is random error which shows how the fitted model is different from the actual model. It reflects the difference between observed and fitted model. Here this model is for only one explanatory variable, but if we have more than one explanatory variables, multiple linear regression model is used. In multiple linear regression more than one factors are considered for affecting the response (Load).

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \quad (2)$$

Above representation is based on matrix dimensions as shown below, and solvable in (3)

$$[Y] = [X\beta] + [\varepsilon] \quad (3)$$

Y : $n \times 1$ vector of observations on study variable or response variable

X : $n \times k$ matrix of n observations each of the k independent variables

$\beta = (\beta_1, \beta_2, \dots, \beta_k)^T$: $k \times 1$ vector of regression coefficients associated with x_1, x_2, \dots, x_n

$\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)^T$: $n \times 1$ vector of random errors

Interceptor term takes first column of X to be $(1, 1, \dots, 1)^T$

Assumptions:

- $E(\varepsilon) = 0$
- $V(\varepsilon) = E(\varepsilon\varepsilon') = \sigma^2 I_n = \begin{bmatrix} \sigma^2 & 0 & 0 \\ : & \sigma^2 & : \\ 0 & \dots & \sigma^2 \end{bmatrix}$
- $\text{Rank}(x) = k$: full column matrix
- $\varepsilon \sim N(0, \sigma^2, I_n)$

If we have the values of vectors for regression coefficients we can predict the load for given observations. To find the values of regression coefficients method of least squares is used. It is based on the principle of maxima/minima. Here minimization of sum of squared errors is considered.

$$\text{minimize } \sum_{i=1}^n \varepsilon_i^2$$

So from (3) we can write,

$$S(\beta) = \sum_{i=1}^n \varepsilon_i^2 = \varepsilon^T \varepsilon = (Y - X\beta)^T (Y - X\beta) \quad (4)$$

$S(\beta)$ is a real valued, convex, differentiable function where the minimum will always exist. According to principle of minima,

$$\frac{\partial S(\beta)}{\partial \beta} = 0$$

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad (5)$$

So values of parameters can be found out by this equation. Variance of residual also needs to be considered so,

$$\text{Var}(\varepsilon_i) = \sigma^2$$

$$\hat{\sigma}^2 = \frac{SS_{\text{res}}}{n - k} = MS_{\text{res}}$$

SS_{res} is sum of squared errors due to residuals

MS_{res} is mean squared error

Once the values of parameters are known fitted model can be obtained.

Fitted model:

$$\hat{Y} = X\hat{\beta} = X(X^T X)^{-1} X^T Y = HY \quad (6)$$

$$H = X(X^T X)^{-1} X^T \quad (7)$$

Residuals (ε) can be obtained from the difference between observed and fitted values.

$$\hat{\varepsilon} = Y - \hat{Y} = (1 - H)Y = \bar{H}Y \quad (8)$$

Total deviation and Sum of squared residual is given as,

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) \quad (9)$$

\bar{y} is average value and \hat{y}_i is estimated value.

$$SS_T = SS_{\text{reg}} + SS_{\text{res}} \quad (10)$$

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (11)$$

SS_{reg} is sum of squared errors due to regression and SS_{res} sum of squared errors due to residuals. To check how good the model is or how well the model fits, coefficient of determination can be used.

$$R^2 = 1 - \frac{SS_{res}}{SS_T} = \frac{SS_{reg}}{SS_T}$$

3. RESULTS AND ANALYSIS

Historical data of Newyork city of 9 years starting from May 2007 to April 2016 is used to train the model. This data had measurement interval of one hour so there are more than seventy eight thousand observations. For taking account of the effect of weather variables, dry bulb temperature and dew point is considered. The purpose of taking lots of information is due to the fact that, if more data is used to build the model, good accuracy in prediction of the model is obtained. This data contains bad data like missing data, measurement errors and sudden spikes, but load forecast needs fine granularity of data. So the data needs to be cleaned or filtered before it is used to make predictions. This is obtained in our previous work [3]. So the filtered data is used here for load forecasting.

Firstly to forecast the load, regression model needs to be built based on available information which is previous year's data of load consumption and temperature data. Software here used for training the model and forecast the data is MATLAB 2018b. In order to build an accurate model, useful predictors are needed. A common technique with temporal predictors requires breaking them into their separate parts so that they can be varied independently of each other. For that three predictors are created, hour, weekday and is weekend or not. The load data itself can be used as a predictor. Load can be used as predictor or not, can be verified by correlation with load. The peaks in the autocorrelation at the 24 and 168 hour, are lagging, so, lagged predictors of 1 and 7 days can be used. So, two predictors are created one as load of prior day and other as prior week. Response variable is load, as load is to be forecasted and explanatory variables are hour, day of week, temperature, dew point, weather it is weekend or not, load of prior day and prior week.

As shown in research method, model is trained using multiple linear regression and least square estimation. Once the model is trained, it is checked how the model will perform on data. Data, for the year 2012 from data set is removed and load forecast for year 2012 is obtained from trained model. To check the effectiveness of forecasting method mean absolute percentage error (MAPE) can be used.

$$MAPE = 100 \times \text{abs} \left(\frac{\text{Actual} - \text{Forecast}}{\text{Actual}} \right) \quad (13)$$

MAPE considering weather data, for training data is 5.11% and for test data or data it had not seen before is 5.15%. MAPE without considering weather data for training data is 5.05% and for test data is 5.06%. So the trained model is nearly 95% accurate in prediction of a day ahead load and method is accurate. The reason behind decrease in MAPE when we do not consider weather data is for day ahead load forecasting because it is not much dependent on weather variable. So when we do not consider temperature and dew point as explanatory variable, MAPE decreases. For further validation of model R^2 can be checked and it is 0.89 and adjusted R^2 is also 0.89, which means model is able to explain nearly 90% of the variation in load. Here T-statistic and P-values are shown in Table 1. Estimated coefficients:

Table 1. T-statistic and P-values

Predictors	Estimate	SE	tStat	P-value
(Intercept)	671.84	9.1257	73.621	0
Hour	4.4814	0.26971	16.616	6.8434e-62
Day of Week	-77.795	0.83435	-93.241	0
is Weekend	-383.48	3.745	-102.4	0
Prior Day	0.69258	0.0021815	317.47	0
Prior Week	0.25824	0.0022122	116.73	0

If the T-stat is greater than 2(>2) or less than 2(<-2) of each variable, it is acceptable. In table all the values of T-stats are greater than 2 or less than 2, meaning that all the variables are statistically significant. P-values indicate that the population parameter is equal to zero. If it is less than 0.1, it indicates a significant regression. As shown in table most p-values are zero and one is also below 0.1, therefore all the repressors are significant. So the trained model is good and can be used to make prediction. Here a day ahead load forecasting is obtained for 25th June on hourly basis by using trained model as given in Table 2. Table 2 displays forecasted load of 25th June 2019 on hourly basis.

Table 2. Day ahead load forecast for 25th June 2019

Date	Load
'25-June-2019 00:00:00'	6372.58781138663
'25-June-2019 01:00:00'	6300.39910537624
'25-June-2019 02:00:00'	6220.51740344386
'25-Apr-2019 03:00:00'	6131.47813413484
'25-Apr-2019 04:00:00'	6039.62451384033
'25-Apr-2019 05:00:00'	5951.29602464256
'25-Apr-2019 06:00:00'	5890.90261494995
'25-Apr-2019 07:00:00'	5883.09608488828
'25-Apr-2019 08:00:00'	5927.78804857792
'25-Apr-2019 09:00:00'	5999.00993029721
'25-Apr-2019 10:00:00'	6090.84866859028
'25-Apr-2019 11:00:00'	6209.37690108279
'25-Apr-2019 12:00:00'	6365.73829277174
'25-Apr-2019 13:00:00'	6557.63245626927
'25-Apr-2019 14:00:00'	6769.75476047530
'25-Apr-2019 15:00:00'	7000.63098695282
'24-Apr-2019 16:00:00'	7234.84272407561
'24-Apr-2019 17:00:00'	7464.44970725805
'24-Apr-2019 18:00:00'	7662.52883584328
'25-Apr-2019 19:00:00'	7789.90505909816
'25-Apr-2019 20:00:00'	7853.00903591586
'25-Apr-2019 21:00:00'	7874.43590892433
'25-Apr-2019 22:00:00'	7858.58727427392
'25-Apr-2019 23:00:00'	7795.30279339892

In Table 2 day ahead load forecasting is obtained and with the help of day ahead load forecast, week ahead load forecast is also obtained. Week ahead load forecast is obtained by considering day ahead load forecast as actual load data and used for next day load forecast. This process is done for whole week. Here forecast of week ahead is as shown in Figure 1. Figure 1 displays, Week days 25th June to 1st July on X-axis and Load in 1000 to 8000 MW on Y-axis. Weekly forecast cannot be obtained directly because forecast model is dependent on data of previous day. So first day ahead load forecast is obtained and it is used as data, then process is repeated and forecast for next day is obtained Figure 1.

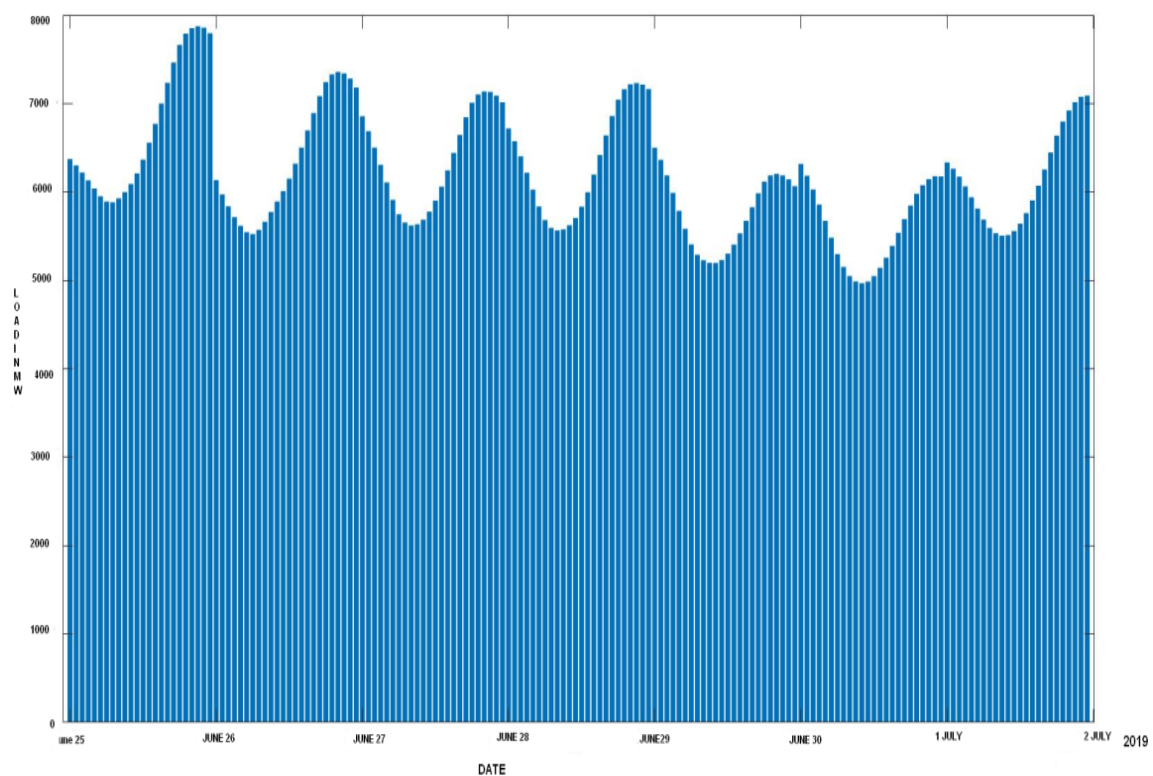


Figure 1. Week ahead load forecast from 25th June to 1st July

4. CONCLUSION

Short Term load forecasting for a day ahead forecast is obtained using multiple linear regression. Use of multiple linear regression is because load is dependent on many variables like historical load and weather variable. Multiple linear regression model for short term load forecasting is easy to develop and MATLAB is used to develop the model. For day ahead load forecast there is no much effect of temperature and due point on load as the mean absolute percentage error with weather variable and without weather variable is nearly same. Model is 95% accurate in prediction. So the day ahead load forecasting is done with the developed model and with the help of day ahead load forecast data, week ahead load forecast is also obtained.

ACKNOWLEDGEMENTS

Thanks to the electrical engineering department, of Faculty of Technology and Engineering M.S.U. of Baroda.

REFERENCES

- [1] Harsh Patel, Mahesh Pandya, Mohan Aware, "Short Term Load Forecasting of Indian System Using Linear Regression and Artificial Neural Network," *5th Nirma University International Conference on Engineering (NUiCONE)*, pp. 1-5, 2015.
- [2] P. Ansarimehr, S. Barghinia, Z. Mirsepassi, H. Habibi, "Identification and Modification of Improper Load Data Used in Short-Term Load Forecasting," *IEEE Russia Power Tech. St.*, pp. 1-5, 2005.
- [3] Dhaval Bhatti, Anuradha Deshpande, "Bad Data detection and Data Filtering in Power System," in *International Journal of Computer Applications*, vol. 182, no. 30, pp. 36-39, December 2018.
- [4] P. Mukhopadhyay, G. Mitra, S. Banerjee, G. Mukherjee, "Electricity load forecasting using fuzzy logic: Short term load forecasting factoring weather parameter," *2017 7th International Conference on Power Systems (ICPS)*, pp. 812-819, 2017.
- [5] M. Ramezani, H. Falaghi, M. -R. Haghifam, G.A. Shahryari, "Short-Term Electric Load Forecasting Using Neural Networks," *EUROCON 2005 - The International Conference on "Computer as a Tool"*, pp. 1525-1528, 2005.
- [6] Francesco Sulla, MattiKoivisto, JanneSeppänen, JukkaTurunen, Liisa C. Haarla,Olof Samuelsson, "Statistical Analysis and Forecasting of Damping in the Nordic Power System," in *IEEE Transactions on Power Systems*, vol. 30, no. 1, pp. 306-315, Jan. 2015.
- [7] I. Moghram, S. Rahman, "Analysis and Evaluation of Five Short-Term Load Forecasting Techniques," in *IEEE Transactions on Power Systems*, vol. 4, no. 4, pp. 1484-1491, Nov. 1989.
- [8] N. Yu and S. Shah and R. Johnson and R. Sherick and M. Hong and K. Loparo, "Big data analytics in power distribution systems," *2015 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, Washington, DC, pp. 1-5, 2015.
- [9] Hossein Akhavan-Hejazi, HamedMohsenian-Rad, "Power systems big data analytics: An assessment of paradigm shift barriers and prospects," *Energy Reports*, vol. 4, pp. 91-100, November 2018.
- [10] T. Hong and S. Fan, "Probabilistic electric load forecasting: A tutorial review," *International Journal of Forecasting*, vol. 32, no. 3, pp. 914-938, July 2016.
- [11] P. Wang, B. Liu, and T. Hong, "Electric load forecasting with recency effect: A big data approach," *International Journal of Forecasting*, vol. 32, no. 3, pp. 585-597, 2016.
- [12] T. Hong, P. Pinson, and S. Fan, "Global energy forecasting competition 2012," *International Journal of Forecasting*, vol. 30, no. 2, pp. 357-363, June 2014.
- [13] Chongqing Kang, Yi Wang, YushengXue, Gang Mu, and Ruijin Liao, "Big Data Analytics in China's Electric Power Industry," in *IEEE Power and Energy Magazine*, vol. 16, no. 3, pp. 54-65, May-June 2018.
- [14] F. L. Quilumba, W.-J. Lee, H. Huang, D. Y. Wang, and R. L. Szabados, "Using smart meter data to improve the accuracy of intraday load forecasting considering customer behavior similarities," in *IEEE Transactions on Smart Grid*, vol. 6, no. 2, pp. 911-918, March 2015.
- [15] N. Ding, Y. B'esanger, and F. Wurtz, "Next-day MV/LV substation load forecaster using time series method," *Electric Power Systems Research*, vol. 119, pp. 345-354, February 2015.
- [16] T. Hong, P. Wang, and L. White, "Weather station selection for electric load forecasting," *International Journal of Forecasting*, vol. 31, no. 2, pp. 286-295, June 2015.
- [17] Hossein Akhavan-Hejazia, Hamed Mohsenian-Rad, "Power systems big data analytics: An assessment of paradigm shift barriers and prospects," *Energy Reports*, vol. 4, pp. 91-100, November 2018.
- [18] Ahmed Yousuf Saber, AfrozAlam, "Short term load forecasting using multiple linear regression for big Data," *IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1-6, 2017.
- [19] Saeed M. Badran, OsasamaB.Abouelatta, "Forecasting Electrical load using ANN combined with Multiple Regression Method,"*The research bulletin of Jordan ACM*, vol. 2, no. 2, pp. 52-58, April 2012.
- [20] Benjamín Zayas, Alfredo Espinosa, Vicky Sanchez, Javier Perez, "Getting ready for data analyticsof electric power distribution systems," *International Journal of Computers*, vol. 2, pp. 179-186, 2017.

- [21] Mandeep Singh, Raman Maini, "Various Electricity Load Forecasting Techniques with pros and cons," *International Journal of Recent Technology and Engineering (IJRTE)*, vol. 8, no. 6, pp. 220-229, March 2020.
- [22] Tao Hang, Min Gul, Mesut E Baran, H Lee Wills, "Modeling and forecasting hourly electric load by multiple linear regression with interactions," *IEEE PES General Meeting*, Providence, RI, pp. 1-8, 2010.
- [23] Trevor Hastie, Robert Tibshirani, Jerome Friedman, "The Elements of Statistical Learning," *Springer*, 2nd edition, 2016.
- [24] S.Saravanan, S.Kannan and C. Thangaraj, "India's Electricity Demand Forecast Using Regression Analysis and Artificial Neural Networks Based on Principal Components," *ICTACT Journal On Soft Computing*, vol. 2, no. 4, pp. 365-370, July 2012.
- [25] Lidong Duan, Dongxiao Niu and Zhihong Gu, "Long and term power load forecasting with multi-regressive Regression analysis," *IITA Proceedings of second International Symposium on intelligent Information Technology Application*, pp 514-518, December 2008. <https://doi.org/10.1109/IITA.2008.397>.
- [26] Heng Shi, Minghao Xu, Ran Li, "Deep Learning for Household Load Forecasting-A Novel Pooling Deep RNN," *IEEE Transactions on Smart Grid*, no. 99, 2017.
- [27] Mosad Alkhathami, "Introduction to Electric Load Forecasting Methods," *Columbia International Publishing Journal of Advanced Electrical and Computer Engineering*, vol. 2, no. 1, pp. 1-12, 2015.
- [28] Issac Adekunle. S, Adeyinka Ajao. A, Chihurumanya Felly. N, Ayokunle Awelewa, "Medium-Term Load Forecasting of Covenant University Using The Regression Analysis Methods," *Journal of Energy Technologies and Policy*, vol. 4, no. 4, pp. 10-16, 2014.
- [29] Enrico Ferrera, Riccardo Tomasi, Xiaolei Hu & Claudio Pastrone, "Evaluation of Short Term Load Forecasting Techniques Applied for Smart Micro Grids," *World academy of science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control & Information Engineering*, vol 9, no. 11, pp. 1942-1947, 2014.

BIOGRAPHIES OF AUTHORS



Dhaval Bhatti Research student 2018-2019, Dept. of Elect. Engg., FTE, MSU of Baroda, Vadodara 390001. Email: dhaval.bhatti4995@gmail.com



Anuradha Deshpande Associate Professor of electrical Engineering, Deptt of Elect Engg, Faculty of Technology and Engineering, M S U of Baroda, Vadodara 390001. She has ME (EPE) in Electrical Engineering. She has book published, many National and International journal papers published in her credit score and reviewer, teacher and guide to students. Email: a.s.deshpande-ee@msubaroda.ac.in